

Supplementary Material: Evolutionary and population genomics of the cavity causing bacteria *Streptococcus mutans*

Omar E Cornejo, Tristan Lefebure, Paulina D Pavinski Bitar, Ping Lang,
Vincent P. Richards, Kirsten Eilertson, David Beighton, Lin Zeng, Sang-Joon Ahn,
Robert A. Burne, Adam C. Siepel, Carlos D. Bustamante, and Michael J. Stanhope

Contents

1 Inference of Core and Pan Genome in <i>Streptococcus mutans</i>	2
2 Selection analyses on core genes and identification of orthologous genes in <i>S. mutans</i> with out-group <i>S. ratti</i> and the set of unique core genes in <i>S. mutans</i>	5
3 Demographic analyses	8
3.1 Assessment of geographic population structure in <i>S. mutans</i>	8
3.2 Demographic reconstruction of <i>S. mutans</i> . Bootstrapped confidence interval estimates of parameters	8
3.3 Simulations with recombination and estimated demographics	10
4 Adaptive value of the unique core genome	10
5 Mapping and assembly pipeline	22

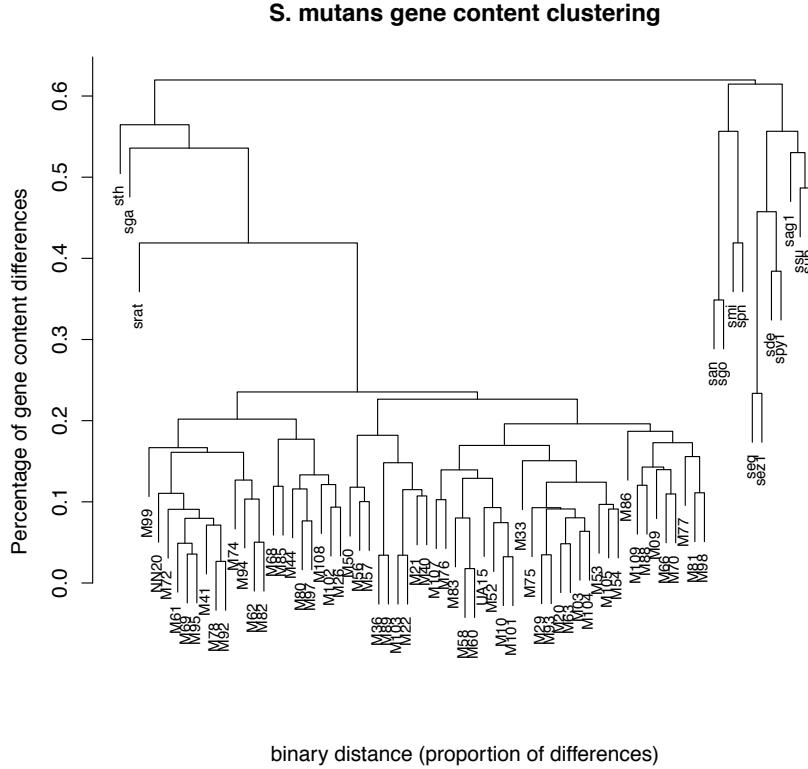


Figure S1: Gene content clustering. The distance used to compute the clustering is the proportion of differences, for example two genomes with 2000 genes each differing by 200 genes will show a dissimilarity of 10%. M69 and M95 show a gene content dissimilarity of 4%

1 Inference of Core and Pan Genome in *Streptococcus mutans*

The clustering of genomes based on gene content is depicted in Figure S1. The core genome refers to genes present in all the isolates of the species (Figure S2). It can be mapped with reads of newly sequenced genomes aligned against reference genomes. To evaluate if we have thoroughly sampled the pan-genome of the species, we performed a permutation test that assessed the decay in the number of new genes observed as the number of samples (genomes) increases (Figure S3). In a similar way to the rarefaction analysis of species composition in a community, the idea is to identify when the decay in gene content, as the number of genomes assessed increases, reaches a plateau. We (and others) have used the same approach to assess pan-genome size for other bacteria species (e.g. [1, 2]). Genes annotated in the de novo assembled genomes served as the primary source of data for this pan-genome curve (Figure S3). Information from mapping, using MAQ, helped refine these estimates. The permutation test suggests that the core genome size estimate is improved by the mapping (1170, *de novo* - versus 1549, *de novo* and mapping, Figure S2). If the estimation is performed allowing each genome to miss one gene, and then mapping is employed to determine if a putatively missing gene in the *de novo* assembly is present, then we get a closer figure to the size of the core genome of the species. For the population genetic analyses involving core genes we kept only 1430 genes which had more than 90% of the length of the gene mapped in all strains. After reciprocal blasting, 82% of the core genes in *S. mutans* have orthologous sequences in *S. ratti*.

The genomes of *Streptococcus criceti* and *Streptococcus macacae* were not included in the initial search for homologous genes because these isolates had not been sequenced at the time. Their later sequencing and inclusion in the work was done with the specific purpose of testing if the unique core sequences present in *S. mutans* were acquired in this species by LGT or lost in the *S. ratti* lineage.

The pan-genome size of *S. mutans* declines significantly after about 20 genomes sequenced, and starts approaching a plateau around 50 genomes, with a total of 3296 genes after 59 genomes sequenced (Figure S3).

On average, only 2 new genes are expected after 59 genomes sequenced, however, in our experience, *de novo* assembly generates a number of small rare genes that are questionable and indeed might be artificial [1]. Therefore, though some very rare genes might still be discovered by additional genome sequencing, this set of 59 genomes can be regarded as a thorough sampling of the *S. mutans* gene repertoire.

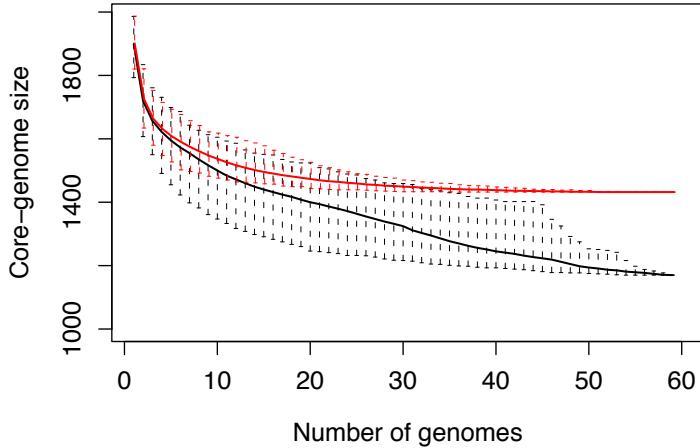


Figure S2: Core genome size estimates using accumulation curves. The input order of the 59 *S. mutans* genomes was randomly permuted 1000 times. Black: *de novo* assembly only, red: *de novo* assembly and allowing one strain to miss one core gene.

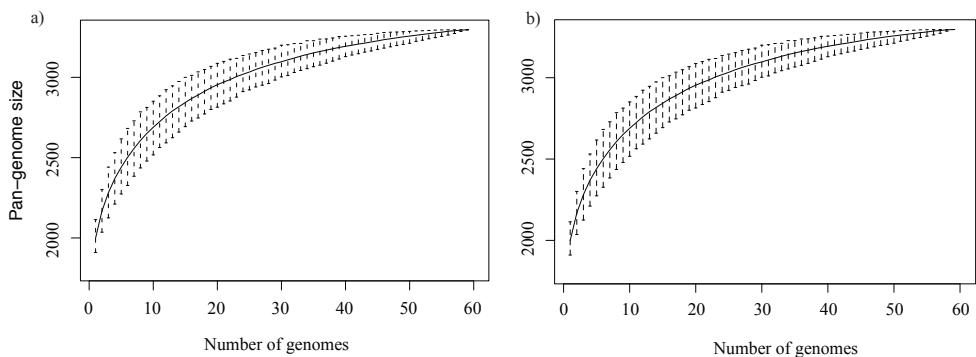


Figure S3: Pan-genome size estimates using accumulation curves. The input order of the 59 *S. mutans* genomes was randomly permuted 1000 times. a) Fit to the model $y = a - cx^{-b}$. b) Fit to the model $y = a + b\ln(x + c)$. In both cases y corresponds to the genome size and x corresponds to the number of genomes. A comparison of the models suggests that there are no significant differences in the fit (ANOVA p-val > 0.05).

Streptococcus mutans. UA159 as reference genome

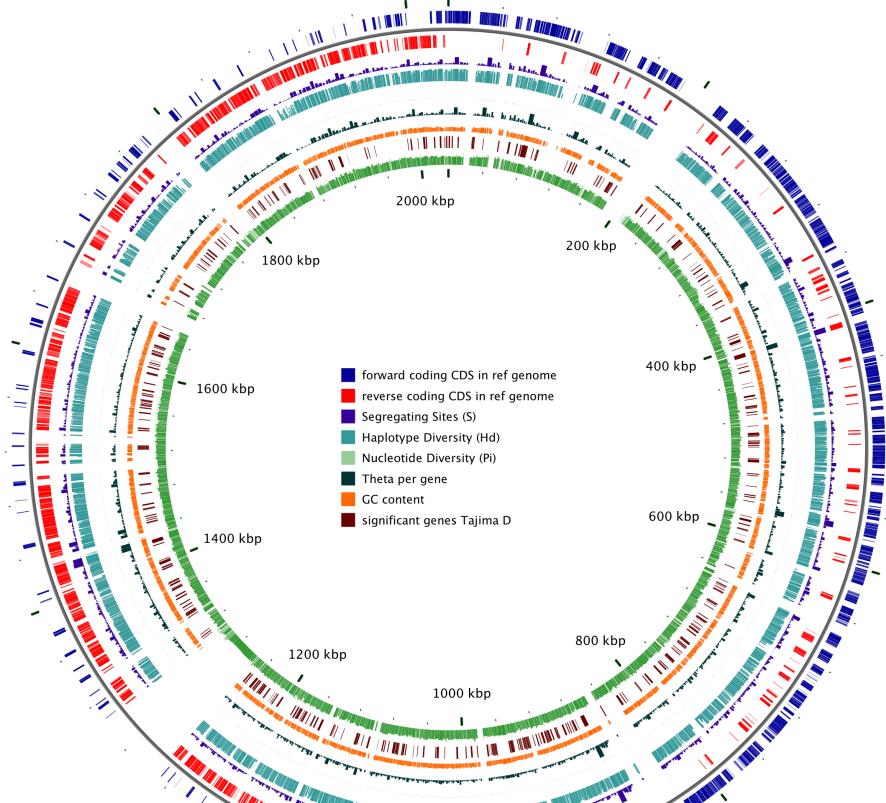


Figure S4: Map of UA159 to show summary statistics for the core genome of *S. mutans*. The outer circles show the core genes coding in the forward (blue) and reverse (red) strands. The inner circles depict information on number of segregating sites per gene (S in purple), haplotype diversity (Hd in aquamarine), nucleotide diversity (π , in light green), and Watterson Theta (θ_W , in dark green), GC content (in orange) and genes with significant Tajima's D (maroon)

Summary Statistics calculated for the core genome of *S. mutans* are presented in Figure S4.

2 Selection analyses on core genes and identification of orthologous genes in *S. mutans* with outgroup *S. ratti* and the set of unique core genes in *S. mutans*

We present in table S1 the genes that show signatures of positive selection, either using simple MK tests or SnIPRE.

Table S1: Genes showing evidence of positive selection

Locus tag (UA159 as reference)	product
SMU.1024	oxaloacetate decarboxylase
SMU.1498	lactose repressor
SMU.1943	leucyl-tRNA synthetase
SMU.365	glutamate synthase (large subunit)
SMU.985	putative beta-glucosidase
SMU.1252	putative glycerate kinase
SMU.1307c	hypothetical protein
SMU.1548c	putative histidine kinase
SMU.1550c	integral membrane protein
SMU.1563	putative cation-transporting P-type ATPase PacL
SMU.1806	putative glycosyltransferase
SMU.864	putative ABC transporter, permease component
SMU.995	ferrichrome ABC transporter permease protein
SMU.998	putative ABC transporter, periplasmic ferrichrome-binding protein

What do we want from this?

- Identify orthologous genes in *S. mutans* and *S. ratti*
- Identify the core genes in *S. mutans* not present in *S. ratti*

We identified 73 *S. mutans* genes that are not present in either one of the three mutans group streptococci and for many of these we identified putative "donors" by sequence similarity (BLAST on the NCBI nr database). Table S2 summarizes information on these genes, including the species that show top BLAST hit similarities.

Table S2: Unique core genes in *S. mutans*

Product	Locus tag (from UA159)	Amino acid length	Top BLAST hit (Species)
hypothetical protein	SMU.1047c	35	NA
hypothetical protein	SMU.1056	39	NA
hypothetical protein	SMU.1131c	87	<i>Listeria sp.</i>
hypothetical protein	SMU.1147c	61	NA
putative DinF, damage-inducible protein; cation efflux pump (multidrug resistance protein)	SMU.121	442	<i>Streptococcus sanguinis</i>
hypothetical protein	SMU.1236c	267	<i>Fusobacterium nucleatum</i>
putative transcriptional regulator	SMU1246c	98	<i>Streptococcus sp.</i>
hypothetical protein	SMU.1267c	120	NA
hypothetical protein	SMU.1393c	275	NA
hypothetical protein	SMU.1395c	42	NA
putative transcriptional regulator	SMU.1409c	288	<i>Streptococcus gallolyticus</i>

Continued on Next Page...

Table S2 – Continued

Product	Locus tag (from UA159)	Amino acid length	Top BLAST hit (Species)
putative reductase	SMU.1410	477	<i>Streptococcus gallolyticus</i>
hypothetical protein	SMU.1456c	42	NA
hypothetical protein	SMU.1502c	81	<i>Streptococcus sanguinis</i>
hypothetical protein	SMU.1504c	108	<i>Ruminococcaceae bacterium</i> and <i>Sebaldella termitidis</i>
putative potassium uptake system protein TrkB	SMU.1561	219	<i>Ethanoligenens harbinense</i>
putative potassium uptake protein TrkA	SMU.1562	217	<i>Ethanoligenens harbinense</i>
hypothetical protein	SMU.1579	66	NA
putative transcriptional regulator	SMU.161	245	<i>Staphylococcus aureus</i>
hypothetical protein	SMU.1616c	156	<i>Streptococcus anginosus</i> , <i>S. sanguinis</i> and <i>S. pyogenes</i>
hypothetical protein	SMU.162c	330	<i>Abiotrophia defectiva</i> and <i>Eubacterium yurii</i>
hypothetical protein	SMU.1641c	66	<i>Streptococcus vestibularis</i> and <i>Streptococcus salivarius</i>
hypothetical protein	SMU.1643c	125	NA
tellurite resistance protein TehB	SMU.1645	293	<i>Streptococcus agalactiae</i>
hypothetical protein	SMU.1648c	81	NA
hypothetical protein	SMU.1655c	46	<i>Leptotrichia buccalis</i> and <i>Leptotrichia hofstadii</i>
hypothetical protein	SMU.18	45	duplication(?)
putative transcriptional regulator	SMU.1805	198	<i>Actinomyces</i> sp. oral taxon and <i>Streptococcus parasanguinis</i>
iron/manganese ABC transporter ATP-binding protein	SMU.182	240	<i>Streptococcus gordonii</i> and <i>Streptococcus sanguinis</i>
putative Mn/Zn ABC transporter	SMU.183	279	<i>Streptococcus cristatus</i> and <i>Streptococcus sanguinis</i>
ABC transporter	SMU.184	306	<i>Streptococcus anginosus</i> , <i>Streptococcus cristatus</i> and <i>Streptococcus sanguinis</i>
hypothetical protein	SMU.185	45	NA
hypothetical protein	SMU.1854	133	<i>Clostridium</i> sp., <i>Streptococcus suis</i> , <i>Streptococcus gordonii</i> and <i>Streptococcus sanguinis</i>
hypothetical protein	SMU.1856c	239	<i>Abiotrophia defectiva</i> , <i>Eubacterium yurii</i> , <i>Solobacterium moorei</i> , and <i>Enterococcus faecalis</i>
hypothetical protein	SMU.1861c	81	<i>Streptococcus sanguinis</i>
hypothetical protein	SMU.1862	67	NA
hypothetical protein	SMU.189	68	NA
putative transcriptional regulator	SMU.1926	191	<i>Lactobacillus casei</i> , <i>Lactobacillus paracasei</i> , <i>Leuconostoc fallax</i>
putative ABC transporter, permease protein	SMU.1928	870	<i>Lactobacillus brevis</i> , <i>Lactococcus lactis</i>
hypothetical protein	SMU.1976c	146	<i>Streptococcus vestibularis</i> , <i>Streptococcus thermophilus</i> and <i>S. sanguinus</i>
hypothetical protein	SMU.2033c	617	NA
hypothetical protein	SMU.2048	50	NA
putative transcriptional regulator	SMU.2058	264	<i>Streptococcus parauberis</i>

Continued on Next Page...

Table S2 – Continued

Product	Locus tag (from UA159)	Amino acid length	Top BLAST hit (Species)	
putative integral membrane protein	SMU.2059c	341	<i>Streptococcus gallolyticus</i> , <i>Streptococcus bovis</i>	
LysR family transcriptional regulator	SMU.2060	291	<i>Streptococcus equinus</i> , <i>Streptococcus bovis</i> and <i>Streptococcus gallolyticus</i>	
hypothetical protein	SMU.2061	217	<i>Streptococcus sanguinis</i> and <i>Streptococcus cristatus</i>	
hypothetical protein	SMU.2090c	50	NA	
hypothetical protein	SMU.2113c	181	<i>Lachnospiraceae oral</i> and <i>Eubacterium cel-lulosolvens</i>	
putative purine-nucleoside phosphorylase	SMU.2126c	253	<i>Streptococcus anginosus</i> and <i>Streptococcus sanguinis</i>	
hypothetical protein	SMU.2136c	58	NA	
hypothetical protein	SMU.31	206	<i>Treponema vincentii</i> , <i>Peptoniphilus sp. oral taxon</i> and <i>Streptococcus bovis</i>	
hypothetical protein	SMU.390	52	NA	
putative (R)-2-hydroxyglutaryl-CoA dehydratase activator-related protein	SMU.438c	1433	<i>Lactococcus lactis</i>	
hypothetical protein	SMU.444	35	NA	
hypothetical protein	SMU.448	126	<i>Streptococcus pyogenes</i>	
hypothetical protein	SMU.451	34	NA	
hypothetical protein	SMU.503c	211	<i>Streptococcus gordonii</i> , <i>Streptococcus vestibularis</i> , <i>Streptococcus salivarius</i>	
hypothetical protein	SMU.529	37	NA	
hypothetical protein	SMU.545	46	NA	
hypothetical protein	SMU.594	51	NA	
hypothetical protein	SMU.622c	170	NA	
hypothetical protein	SMU.631	252	<i>Streptococcus anginosus</i> , <i>Streptococcus oralis</i>	
hypothetical protein	SMU.68	25	NA	
hypothetical protein	SMU.722	59	NA	
hypothetical protein	SMU.748	58	NA	
fructan hydrolase; fructosidase; FruB	exo-beta-D-	SMU.79	519	<i>Lautropia mirabilis</i> , <i>Leuconostoc mesenteroides</i>
hypothetical protein	SMU.847c	134	<i>Streptococcus agalactiae</i>	
hypothetical protein	SMU.851	163	<i>Streptococcus porcinus</i> and <i>Faecalibacterium prausnitzii</i>	
hypothetical protein	SMU.914c	127	<i>Staphylococcus aureus</i> and <i>Pediococcus acidilactici</i>	
hypothetical protein	SMU.958	106	<i>Pediculus humanus corporis</i>	
hypothetical protein	SMU.959c	84	NA	
hypothetical protein	SMU.986c	171	<i>Streptococcus oralis</i> and <i>Streptococcus infantis</i>	
ferrichrome ABC transporter permease protein	SMU.996	323	<i>Streptococcus agalactiae</i>	

3 Demographic analyses

3.1 Assessment of geographic population structure in *S. mutans*

The principal component analysis was performed in R, using the total synonymous SNP data matrix generated with Polydnds, as well as with a pruned data matrix considering only SNPs with frequencies greater than 5%. Both produced similar results (Figure S5).

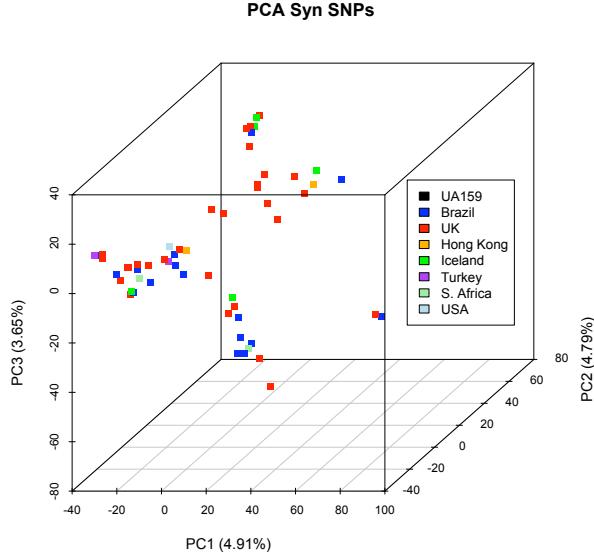


Figure S5: Principal Component Analysis on *S. mutans* synonymous SNPs. Color coding corresponds to the different geographic locations. The percentage of variation explained by each PC is shown within parentheses.

3.2 Demographic reconstruction of *S. mutans*. Bootstrapped confidence interval estimates of parameters

We evaluated 5 different demographic scenarios using $\delta\alpha\delta i$ [3]. These scenarios are schematically depicted in Figure S6. $\delta\alpha\delta i$ [3] implements a Wright-Fisher model *via* a diffusion approximation that numerically computes the expected site frequency spectrum under the assessed demographic scenario. The method provides for the selection of the best combination of parameters that maximize the likelihood of observing the data. The best model (demographic scenario) explaining the data is selected based on the Akaike Information Criteria (AIC). After selecting the exponential growth model, the absolute fit of the expected site frequency spectrum to the observed site frequency spectrum was evaluated employing a Kolmogorov-Smirnov comparison of distributions. The likelihood surface of the parameter search was evaluated to make sure that it did not fall into local maxima. Models were searched with different initial parameter values and evaluated if they all converged to the same maximum. We used a grid size of 90. The likelihood surface in Figure 1 was obtained by finding the likelihood value for each combination of parameters in a grid search by brute force: all combinations of parameter ν (from 1 to 20 every 1) and parameter τ (from 0.1 to 10 every 0.1). We estimated the confidence intervals for the parameters by bootstrapping the synonymous data matrix and fitting the model. The sum vector of the minor allelic frequency for the synonymous matrix was re-sampled with replacement 1000 times, creating an equal number of resampled vectors with the same dimensions as the original one. From each of these vectors, site frequency spectra were constructed and ML values of the parameters were inferred under the exponential growth model using $\delta\alpha\delta i$.

Histograms depicting 95% confidence intervals for the parameters with conservative estimates of time are shown in Figure S7.

Estimations performed with $\delta\alpha\delta i$ recover two main parameters under the exponential growth model. The first is ν , the ratio of current to ancestral population size; and the second is τ , the time since the change in demographic, representing in our case the start of the demographic expansion. τ is equal to time (T) scaled by $2N_a$ (which is the ancestral population size). With $\delta\alpha\delta i$, we can also obtain an optimized value of the parameter $\theta = 2N_e\mu$. Because the estimated parameters are scaled by θ , to obtain estimates of time in years it is necessary to re-scale the values obtained with $\delta\alpha\delta i$ in the following manner:

the optimized θ obtained from $\delta\alpha\delta i$ is expressed in terms of N_a .

$$\theta = 2N_a\mu \quad (1)$$

where μ is the mutation rate for the total number of synonymous sites examined. As mentioned in the Materials and Methods, experimental estimates of the spontaneous rates of mutation are available. These estimates are expressed in terms of substitutions per site per replication (generation in asexually reproducing organisms), and have been estimated via fluctuation tests. We take as a range, the estimates reported by Drake [4] which suggest mutation rates of $\mu_{bp} = \{4.08 \times 10^{-10}$ and $6.93 \times 10^{-10}\}$ changes per site per generation. These estimates, although arising from *Escherichia coli* and *Salmonella enterica*, are suitable for our situation because experimental determinations of spontaneous mutation rates for antibiotic resistance markers (such as spectinomycin, streptomycin and rifampicin) in *S. mutans* are not different from those estimated in other species of bacteria like *E. coli*, *Streptococcus pneumoniae*, *Streptococcus mitis*, etc ([5–7]). There is no clear calibration point (in time) used to perform estimates of the mutation (i.e. substitution) rate in *Streptococcus mutans* based on a comparison with *S. ratti* or *S. criceti*. Furthermore, Gibbons [8], and Berkowitz [9] provide estimates of generation times (number of division cycles) for *S. mutans* ranging from 2 - 4 divisions per day, equaling 730 - 1460 divisions per year. Based on this, we estimated a minimum and maximum mutation rate per synonymous genome, by multiplying the maximum (or minimum) mutation rate by the largest (or lowest) number of generations per year and the number of potential synonymous sites considered in the inference of synonymous site frequency spectrum. Our figures suggest a synonymous genome mutation rate (μ_{sg}) ranging from $\mu_{min} = 0.112 \text{ subs/syn_genome/year}$ to $\mu_{max} = 0.379 \text{ subs/syn_genome/year}$.

We then consider the following:

$$\tau = \frac{T}{2N_a} \quad (2)$$

$$T = \tau 2N_a \quad (3)$$

$$(4)$$

from θ we can infer:

$$2N_a = \frac{\theta}{\mu} \quad (5)$$

leaving:

$$T = \tau \frac{\theta}{\mu} \quad (6)$$

We used the optimized estimates of the mutation parameter θ obtained from fitting the model, the 95% lower and upper estimates of τ obtained from fitting the bootstrapped data sets, and the minimum and maximum mutation rates per year, to obtain the re-scaled 95% confidence intervals for our estimates of the time since start of the expansion. These confidence intervals also accommodate the uncertainties in mutation rate estimates and generation times. The result is a confidence interval ranging from 3,268.42 ya to 14,344.26 ya.

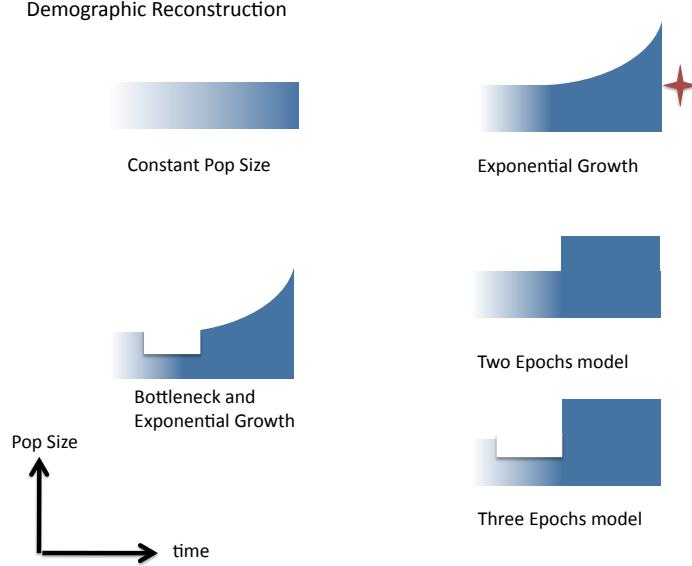


Figure S6: Schematic representations of the five demographic models assessed by fitting the observed site frequency spectrum to simulations in a maximum likelihood framework. A red star depicts the scenario with the best model fit. (see Table 1 in manuscript)

3.3 Simulations with recombination and estimated demographics

As mentioned in the main body of the manuscript, the methods employed in this work are sensitive to genetic linkage among markers. The expectation of the sfs under linkage is not strongly affected, but the variance is affected. In Figure S8 we present the results from the analyses using ClonalFrame. The left panel of the figure presents the behavior of the chain with generations post-burnin; it is clear that the chains were well mixed. The right panel presents the values of estimated ρ/θ every 100 generations of the chain. Our analyses suggest that the results obtained previously with mlst data are consistent with the results obtained with 600 loci, with a relative contribution of recombination over mutation to variation of approximately 6. Credibility intervals in three separate sets show similar patterns suggesting that this value could be between 3 and 19.

We carried out simulations in ms [10] under gene conversion ($r/\mu \sim 6$), under the same demographic scenario as the one fit to the observed data. The mean site frequency spectrum from 600 simulations and the observed site frequency spectrum present a similar pattern (Figure S9). Then, we estimated parameters ν (ratio of $N_{current}/N_{ancestral}$) and τ (scaled time since start of the expansion) on each one of the simulated sets in $\delta\alpha_i$. Our simulations show a larger variation in the estimated parameters for the simulations (Figure S10), when compared to the classical bootstrap estimates, which could be the result of linkage and a small genome size. Our maximum likelihood estimates of the parameters in the observed data set lies within the confidence interval of the parametric bootstrap.

4 Adaptive value of the unique core genome

We present in table S3 detailed information about the genes identified as part of the unique core genome of *Streptococcus mutans*. The locus tag is preceded by information about the possible role in which the protein products of these genes have been (or could potentially be) implicated.

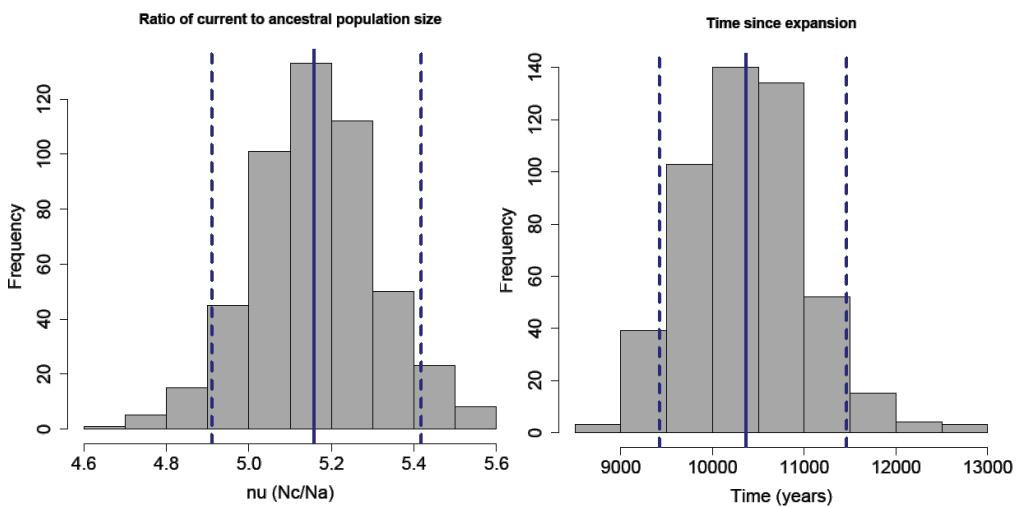


Figure S7: Bootstrapped estimates of the ratio of current to ancestral population size (ν), shown in the left panel and the time since the start of the expansion for the conservative scenario (in years), shown in the right panel. For both figures, the solid line corresponds to the median of the bootstrap estimates and the dashed lines correspond to the limits of the 95% confidence intervals.

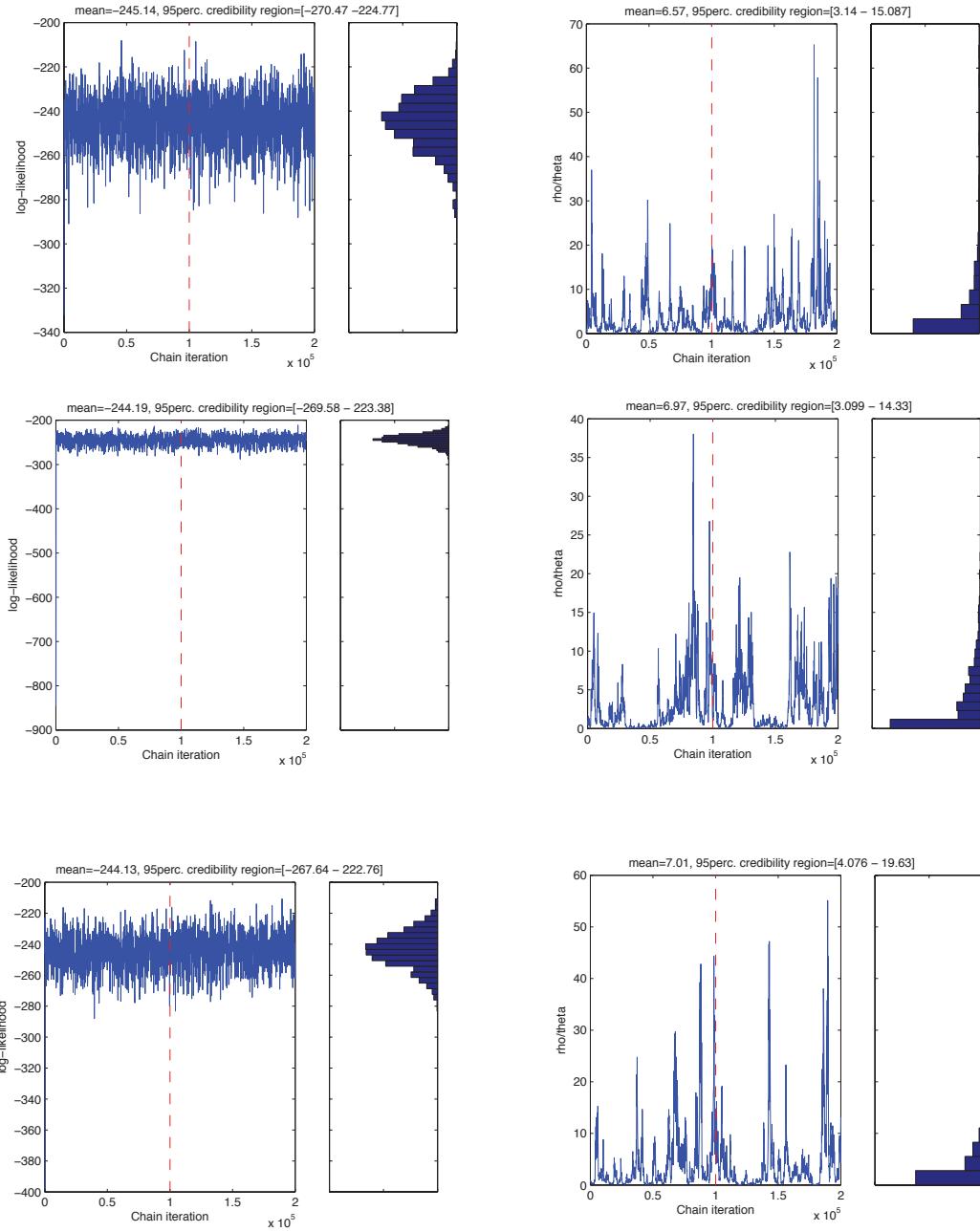


Figure S8: The three panels show results of three sets of simulations on 600 regions (500 bp each) with Clonal Frame. On the left it is shown the mixing of the chains (Log of likelihood every 100 steps). On the right it is shown the estimates of ρ/θ for each set.

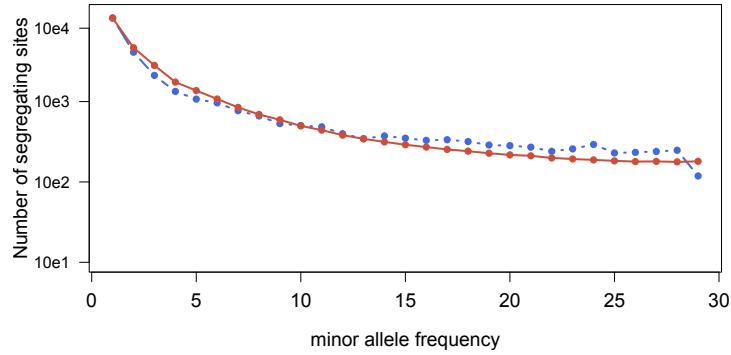


Figure S9: Observed folded site frequency spectrum (blue) and mean site frequency spectrum from 600 ms simulations (red)

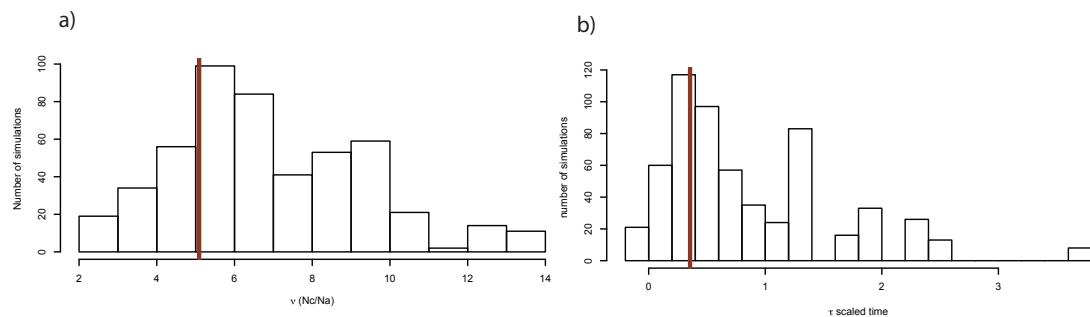


Figure S10: a) Distribution of estimated ν (ratio of N_c/N_a) on the simulated data set. b) Distribution of estimated τ (scaled time since the start of the expansion). In both histograms, the red lines correspond to the estimated parameters from the observed data set.

Table S3: Putative roles of *S. mutans* unique core genes in oxidative stress and regulation

Locus tag #	Description	Secreted	Possible Role	Context
SMU.1047c	Small Peptide	No	Oxidative Stress	Linked to relQ operon and rex (Smu.1053), both affect oxidative stress tolerance [11, 12]
SMU.1056	Small Peptide	No	Acid Oxidative Stress	Linked to radC (DNA repair) and sat operon (protein secretion, acid oxidative stress tolerance) [13]
SMU.1131c	CiaX	Yes	Signaling Stress	Part of CiaRH TCS1 signaling complex affecting stress tolerance and genetic competence [14, 15]
SMU.1147c	Small Peptide	Possibly	Signaling Regulation	Co-transcribed with TCS [Burne personal observations]
SMU.1236c	Amido-hydrolase	No	Signaling Regulation and stress	Linked to GTPase (5') and DNA Repair (3') [16, 17]
SMU.1246c	TetR-family DNA Binding Protein	No	Regulation, Possible CHO metabolism	Hydrolase (5'), Endolase (3') [18]
SMU.1393c	Hyp Small Peptides	No/Possibly	Growth Control, Signaling	In operon with LepA elongation factor [Burne unpublished observation] ^a
SMU.1395	Small Peptide	No	Stress	Possibly co-transcribed with mutX mutator gene [Burne unpublished observation] ^a
SMU.1456	Peptides	No	Unknown	Contiguous, possibly co-transcribed, linked to exomerase (3O) [19]
SMU.1502c and Smu1504c	Potassium transport (trkA)	Yes	Homeostasis Stress tolerance	In operon with P-type ATPase [20–22]
SMU.1562	Small protein	No	Unknown	Co-transcribed with biotin ligase [Burne unpublished observation] ^a
SMU.1579	Hypothetical	No	Stress	Co-transcribed with PadR-regulator (3') and MerR regulator (5'), which regulate genes in response to xenobiotics and other stressors [Burne unpublished observation] ^a
Smu.1641	CsbD-like	Small Peptide	Stress Signaling	CsbD is a gene stress protein of unknown function [Burne unpublished observation] ^a
Smu.1645	tehB D Tellurite resistance	No	Oxidative stress	Linkage to 1648 [23, 24]

Continued on Next Page...

Table S3 – Continued

Locus tag #	Description	Secreted	Possible Role	Context
Smu.1648	Peptide (70aa)	Possibly membrane	Regulation Stress	Co-transcribed with ArsR-like DNA binding protein (HTH) and predicted hemolysin. Linked to tehB (tellurite resistance) and exoA (SmuX D low pH-inducible DNA repair enzyme [Burne unpublished observation])
Smu.1655	Peptide (46aa)	No	Unknown	In serine biosynthetic operon [Burne unpublished observation] ^a
SMU.1730c	aryalkylamine N-acetyltransferase	No	Folate biosynthesis, cell wall metabolism	Possible folate biosynthesis, possibly in operon with peptidoglycan ligase [25]
Smu.18	Peptide	No	Signaling	Co-transcribed with amino acid permease [Burne unpublished observation] ^a
Smu.185	Peptide	Yes	Signaling Oxidative Stress	In middle of slo operon immediately 3' to ABC transporter SloR mediates stress tolerance and Mn ion homeostasis [26]
Smu.1854	HdrR	No	Signaling Bacteriocins	DNA binding protein in HdrRM regulatory complex D controls gene expression at high cell density [27,28]
Smu.1861	Divergently transcribed peptides	Yes	Signaling Regulation	1st gene in possible operon with ribosomal proteins and single-stranded binding protein (Smu.1862). mutY is downstream following a large IGR (IGR1461) (sRNA?) [Burne unpublished observation] ^a
Smu.1862				Immediately 5' to and divergently transcribed from hasp33-like molecular chaperone involved in oxidative stress and redox sensing [Burne unpublished observation] ^a
Smu.189	Peptide	Yes	Signaling Stress?	Co-transcribed with sapR TCS and 5' to comDEC signaling system involved in competence, bacteriocin production, stress tolerance and biofilm formation [29]
Smu.1918	dedA (Selenite Resistance)	Membrane	Stress Signaling Competence	DS of ackA (acetylphosphate metabolism) and co-transcribed with Qabtitive infection protein O and membrane protein [Burne unpublished observation] ^a
Smu.1976	Transcriptional Regulator	No	Oxidative metabolism Signaling	

Continued on Next Page...

Table S3 – Continued

Locus tag #	Description	Secreted	Possible Role	Context
Smu.2048	Peptide	Maybe	Stress Signal	Divergently transcribed upstream of ptaA (Maltose PTS EII component), relA, and exonuclease phosphatase [Burne unpublished observation] ^a
Smu.2090	Peptide (50aa)	Probably	Signaling Stress Competence	Between hexA and hexB DNA repair enzymes ^b possibly co-transcribed with hexAB and recA, ruvA and cinA (competence damage-inducible protein) [Burne unpublished observation] ^a
Smu.2113	Protein (phosphoglucomutase domain)	No	CHO binding or metabolism	In operon with and 5O to gbpA (glucan binding protein) and downstream of MerR-family regulator [observed in this study] and [30]
Smu.2126	Phosphorylase (nucleoside phosphorylase)	No	Regulation Oxidative metabolism	5O to TCS oxidoreductase [Burne unpublished observation] ^a
Smu.2136	Peptide (58aa)	No	Signal Growth regulation	In operon with ribosomal proteins, dnaC, gldA [Burne unpublished observation] ^a
Smu.28	ComA-like transporter	ABC Membrane	Transport Growth	In possibly operon with acyl carrier protein and purine biosynthetic genes [31, 32]
Smu.31	Unknown	No	Nucleotide metabolism	Co-transcribed in purine operon [Burne unpublished observation] ^a
Smu.390	Peptide	No	Signaling Catabolism of salivary proteins Oxidative stress	In operon with aa uptake, endopeptidases D 1st gene is MarR-like transcriptional regulator [Burne unpublished observation]
Smu-438	Glutamate biosynth	AA	Nutritional Oxidative stress	Linked to transposase and MarR operon with signaling peptides [this work, Burne unpublished observation]
Smu.444	Peptide	No	Stress Signaling	In operon with MarR-type DNA binding protein; downstream of TetR regulator and hypotheticals [this work and Burne unpublished observation]
Smu.451	Secreted peptide	Yes	Signal Envelope integrity	In operon upstream of methyltransferase and cell wall metabolism enzymes [this work, and Burne unpublished observation]

Continued on Next Page...

Table S3 – Continued

Locus tag #	Description	Secreted	Possible Role	Context
Smu.529	Peptide	Possibly	Envelope Stress	Linked to MerR regulator and diaminopimelate metabolism (cell wall) [this work, Burne unpublished observation]
Smu.545	Peptide	No	Signal Oxidative stress CHO metabolism	Co-transcribed with <i>Dpr</i> , glucose kinase, GTP binding protein (<i>tppA</i>) D followed by cell wall enzymes [this work, Burne unpublished observation]
Smu.622	Hypothetical	Transmembrane	Saliva polysaccharide Metabolism	In operon with polysaccharide deacetylase and protein with both exinuclease and methyltransferase domains [this work, Burne unpublished observation]
Smu 631	Hypothetical	No	Oxidative stress	In operon MarR transcriptional regulator and thioesterase. Immediately upstream of queA (queosine modification of RNA) [this work, Burne unpublished observation]
Smu.642	Small (93aa)	Protein	No	Oxidative metabolism Signaling Competence
Smu.68	Peptide	Possibly	Signaling Regulation Envelope	In operon with protein tyrosine phosphatase and secreted acyltransferase 3 (May O-acetylate wall associated with resistance to lysozyme in <i>Staphylococcus aureus</i> [33])
Smu.722	Peptide	No	Regulation Homeostasis	Possibly in operon with Ca ⁺⁺ ATPase and Ca ⁺⁺ binding protein [Burne unpublished observation]
Smu.748	Peptide	Yes	Signal Stress	Divergently transcribed with overlapping promoters for permease and FetB export pump [Burne unpublished observation]
Smu.847	Protein (134aa)	Membrane	Signal Growth regulation	Encoded antisense overlapping with ribosomal proteins [Burne unpublished observation]
Smu.851	NUDIX hydrolase	No	Regulation CHO metabolism	Hydrolyzes nucleotide diphosphates linked to R group e.g. UDP-sugar linked to acetyltransfer (GNAT) [34]

Continued on Next Page...

Table S3 – Continued

Locus tag #	Description	Secreted	Possible Role	Context
Smu.914	Hypothetical	Membrane	Signaling Stress resistance	Downstream of GTP cyclohydrolase, "Radical SAM" catalyzing unusual methylations and Aluminum-resistance ATPase [Burne unpublished observation]
Smu.958	Peptides	Yes	Signaling Growth Control	Antisense and overlapping with rpl proteins [Burne unpublished observation]a

(from UA159)

a (corroborated at Orlagene web browser)

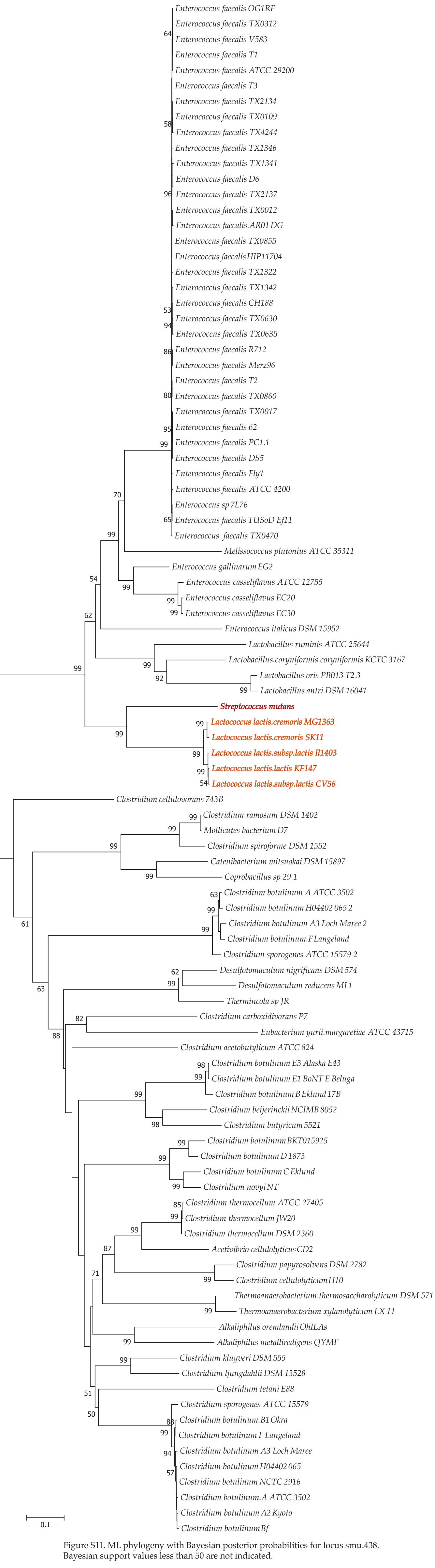


Figure S11. ML phylogeny with Bayesian posterior probabilities for locus smu.438. Bayesian support values less than 50 are not indicated.

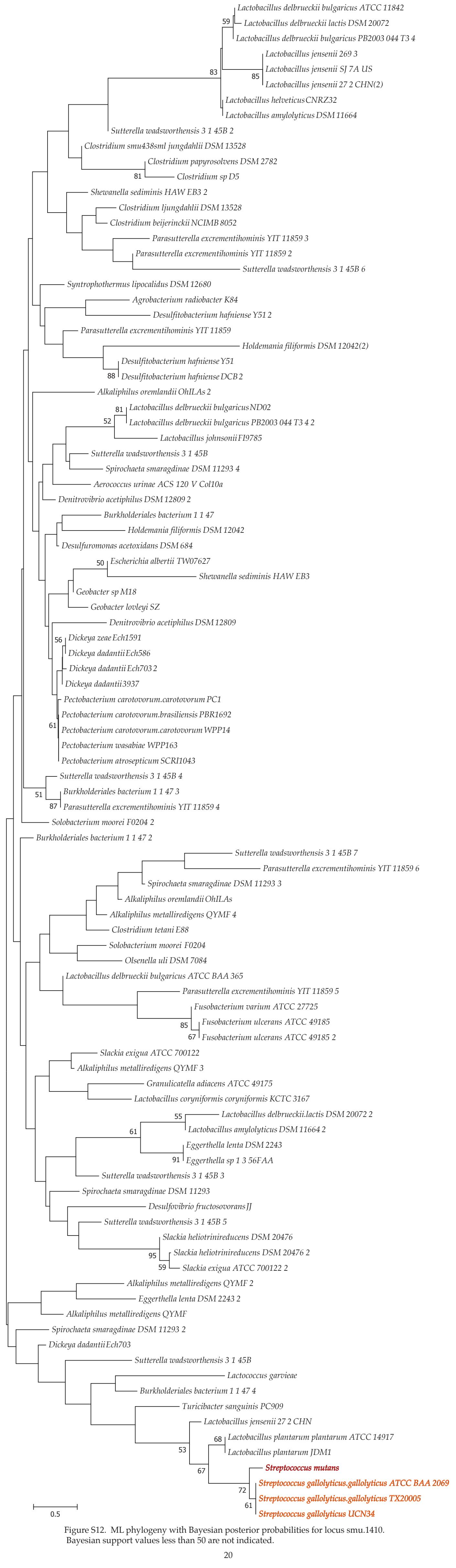


Figure S12. ML phylogeny with Bayesian posterior probabilities for locus smu.1410. Bayesian support values less than 50 are not indicated.

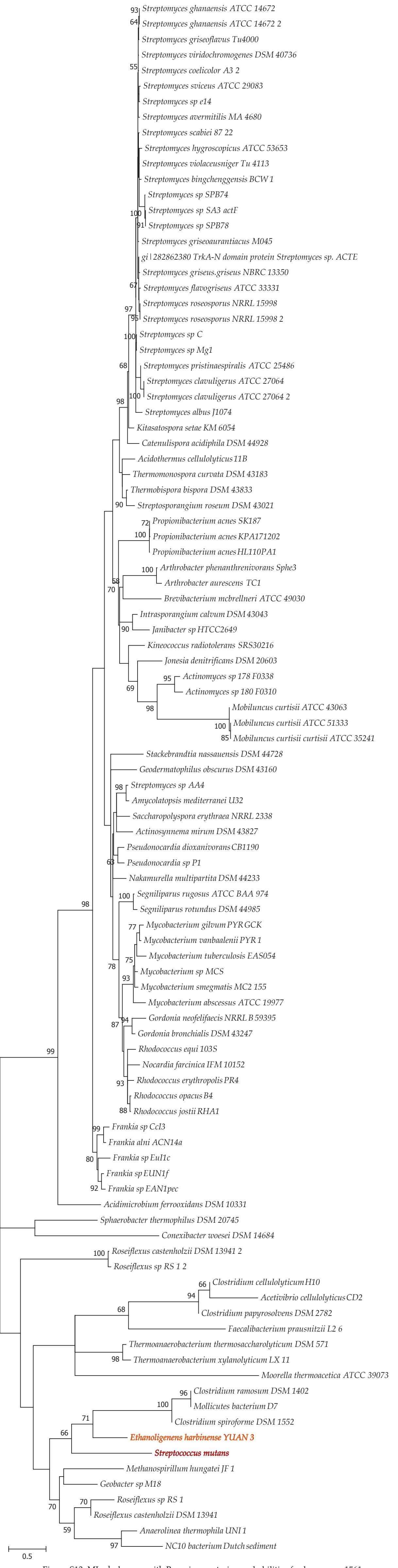


Figure S13. ML phylogeny with Bayesian posterior probabilities for locus smu.1561. Bayesian support values less than 50 are not indicated.

5 Mapping and assembly pipeline

Information on each of the sequenced strains is shown in Table S4.

Table S4: Strains sequenced in this study

Lab ID	Strain Name	Genbank Accession	Country of Origin	MLST	Serotype
SMU 10	8ID3	AHRB00000000	Brazil	ST 1	c
SMU 101	U2B	AGWE00000000	Turkey	ST 49	c
SMU 102	S1B	AHRC00000000	Turkey	ST 55	c
SMU 103	SA38	AHRD00000000	South Africa	ST 113	c
SMU 104	SA41	AHRE00000000	South Africa	ST 114	c
SMU 105	SF12	AHRF00000000	USA	ST 57	c
SMU 107	R221	AHRG00000000	Brazil	NA	c
SMU 108	M230	AHRH00000000	Brazil	NA	e
SMU 109	OMZ175	AHRI00000000	unknown	NA	f
SMU 20	15JP3	AHRJ00000000	Brazil	ST 12	NA
SMU 21	1SM1	AHRK00000000	Brazil	ST 13	e
SMU 22	4SM1	AHRL00000000	Brazil	ST 10	c
SMU 26	3SN1	AHRM00000000	Brazil	ST 16	e
SMU 29	2ST1	AHRN00000000	Brazil	ST 4	e
SMU 3	11A1	AHRO00000000	Brazil	ST 3	c
SMU 33	11SSST2	AHRP00000000	Brazil	ST 18	c
SMU 36	4VF1	AHQ00000000	Brazil	ST 22	NA
SMU 40	15VF2	AHRR00000000	Brazil	ST 25	e
SMU 41	2VS1	AHRS00000000	Brazil	ST 11	c
SMU 44	11VS1	AHRT00000000	Brazil	ST 28	e
SMU 50	5SM3	AHRU00000000	Brazil	ST 14	c
SMU 52	NFSM2	AHRV00000000	UK	ST 2	c
SMU 53	NVAB	AHRW00000000	USA	ST 3	c
SMU 54	A9	AHRX00000000	UK	ST 18	c
SMU 56	N29	AHRY00000000	UK	ST 11	c
SMU 57	NMT4863	AHRZ00000000	Japan	ST 12	c
SMU 58	A19	AHSA00000000	UK	ST 14	c
SMU 60	U138	AHSB00000000	UK	ST 4	c
SMU 61	G123	AHSC00000000	UK	ST 19	c
SMU 62	M21	AHSD00000000	UK	ST 22	c
SMU 63	T4	AHSE00000000	UK	ST 23	c
SMU 66	N34	AHSF00000000	UK	ST 23	c
SMU 68	NFSM1	AHSG00000000	UK	ST 32	c
SMU 69	NLML4	AHSH00000000	UK	ST 37	e
SMU 70	NLML5	AHSI00000000	UK	SST 38	c
SMU 72	NLML9	AHSJ00000000	UK	ST 40	c
SMU 74	M2A	AHSK00000000	UK	ST 42	c
SMU 75	N3209	AHSL00000000	UK	ST 44	c
SMU 76	N66	AHSM00000000	UK	ST 45	c
SMU 77	NV1996	AHSN00000000	USA	ST 47	c
SMU 78	W6	AHSO00000000	UK	ST 50	e
SMU 80	SF1	AHSP00000000	USA	ST 56	c
SMU 81	SF14	AHSQ00000000	USA	ST 58	c
SMU 82	SM6	AHSR00000000	Hong Kong	ST 62	c

Continued on Next Page...

Table S4 – Continued

Lab ID	Strain Name	Genbank Accession	Country of Origin	MLST	Serotype
SMU 83	ST1	AHSS00000000	UK	ST 63	c
SMU 85	ST6	AHST00000000	UK	ST 67	c
SMU 86	U2A	AHSU00000000	Turkey	ST 69	e
SMU 88	NLML8	AHSV00000000	UK	ST 101	c
SMU 89	NLML1	AHSW00000000	UK	ST 103	e
SMU 9	1ID3	AHSX00000000	Brazil	ST 7	e
SMU 92	14D	AHSY00000000	Iceland	ST 71	e
SMU 93	21	AHSZ00000000	Iceland	ST 73	e
SMU 94	66-2A	AHTA00000000	Iceland	ST 77	c
SMU 95	B	AHTB00000000	Iceland	ST 84	c
SMU 97	SM4	AHTC00000000	Hong Kong	ST 61	c
SMU 98	SM1	AHTD00000000	Hong Kong	ST 116	c
SMU 99	24	AHTE00000000	Iceland	ST 6	c

The pipeline followed for the mapping, assembly and annotation of the genes in *S. mutans* is presented in Figure S14. In figure S15 we present a schematic of the pipeline employed to assembly and annotation of the genes in *S. ratti*.

Mapping, Assembly and Annotation Pipeline

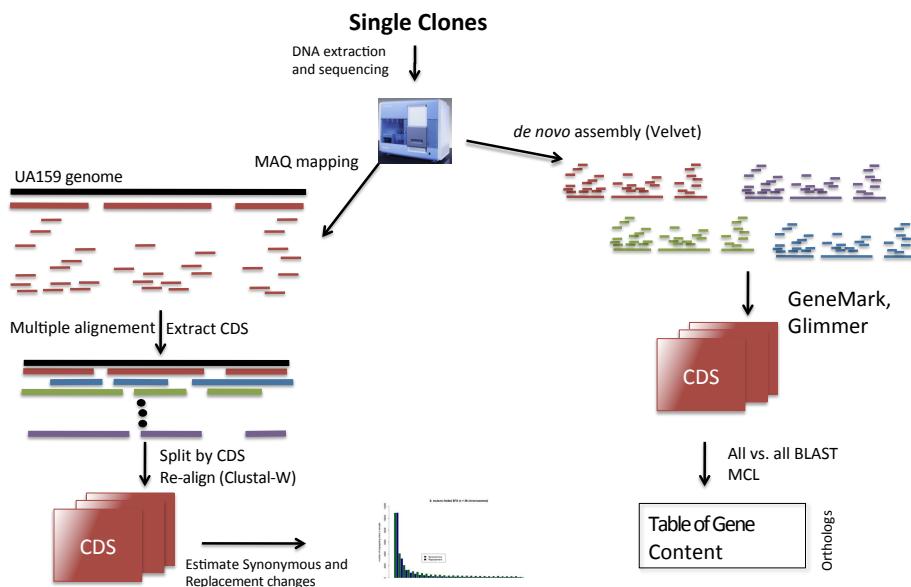


Figure S14: Pipeline for mapping the reads of newly sequenced isolates of *S. mutans* against the reference genomes and realignment of CDS to correct for misalignment of the coding regions (left). *De novo* assembly of the isolates, annotation of CDS and generation of gene content tables per newly sequenced genome (right)

An orthology search was performed using the 57 *de novo* assembled genomes, the reference genomes UA159 and NN2025, and 13 other streptococci genomes available in GenBank (Table S5). An all-versus-all BLAST

search was performed and orthologs delimited using orthoMCL2 [35].

Sequencing the closest related species: *Streptococcus ratti*

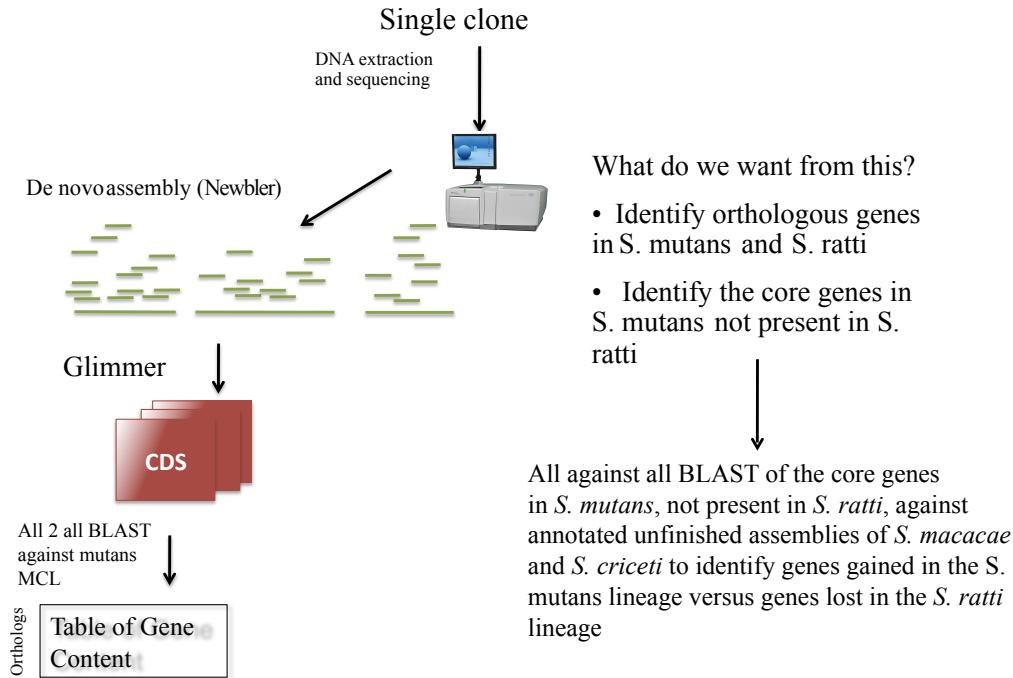


Figure S15: Pipeline for assembling *S. ratti* reads from 454 sequencing, annotation of genes and identification of orthologous genes between *S. ratti* and *S. mutans* via OrthoMCL (left). Additional comparisons of the “unique” core genes in *S. mutans* with two other mutans group streptococci - *S. macacae* and *S. criceti*, provided a means to distinguish which genes were lost on the *S. ratti* lineage and which were gained in *S. mutans* via LGT.

Table S5: NCBI Refseq data used in the study

Taxa	locus tag	short name	refseq accr
<i>Streptococcus agalactiae</i> 2603V/R	sag_2603	sag1	NC_004116
<i>Streptococcus dysgalactiae equisimilis</i> GGS 124	sde_GGS	sde	NC_012891
<i>Streptococcus equi</i> 4047	seq_4047	seq	NC_012471
<i>Streptococcus equi zooepidemicus</i>	sez	sez1	NC_012470
<i>Streptococcus gordonii</i> str Challis substr CH1	sgo	sgo	NC_009785
<i>Streptococcus mutans</i> UA159	smu	smu	NC_004350
<i>Streptococcus mutans</i> NN2025	SmuNN2025	SmuNN2025	NC_013928
<i>Streptococcus pneumoniae</i> TIGR4	spn	spn	NC_003028
<i>Streptococcus pyogenes</i> M1 GAS	spy_M1	spy1	NC_002737
<i>Streptococcus sanguinis</i> SK36	san	san	NC_009009
<i>Streptococcus suis</i> SC84	ssu	ssu	NC_012924
<i>Streptococcus thermophilus</i> LMG 18311	sth	sth	NC_006448
<i>Streptococcus uberis</i> 0140J	sub	sub	NC_012004
<i>Streptococcus gallolyticus</i> UCN34 uid46061	GALLO	sga	NC_013798
<i>Streptococcus mitis</i> B6 uid46097	smi	smi	NC_013853

References

- [1] Lefébure, T., Bitar, P. D. P., Suzuki, H. & Stanhope, M. J. Evolutionary dynamics of complete *Campylobacter* pan-genomes and the bacterial species concept. *Genome Biol Evol* **2**, 646–55 (2010).
- [2] Lefébure, T. & Stanhope, M. J. Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. *Genome Biology* **8**, R71 (2007).
- [3] Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics* **5**, e1000695 (2009).
- [4] Drake, J. W. A constant rate of spontaneous mutation in DNA-based microbes. *Proc Natl Acad Sci U S A* **88**, 7160–4 (1991).
- [5] Perry, D. & Kuramitsu, H. K. Genetic transformation of *Streptococcus mutans*. *Infection and Immunity* **32**, 1295–7 (1981).
- [6] Cornejo, O. E., Rozen, D. E., May, R. M. & Levin, B. R. Oscillations in continuous culture populations of *Streptococcus pneumoniae*: population dynamics and the evolution of clonal suicide. *Proceedings of the Royal Society, Biological Sciences (B-Series)* **276**, 999–1008 (2009).
- [7] Cornejo, O. E. *Population Dynamics and Population Genetics of Recombination in Bacteria*. Ph.D. thesis, Emory University, Atlanta, USA (2009).
- [8] Gibbons, R J, R. J. Bacteriology of Dental Caries. *J Dent Res* **43**, Suppl:1021–8 (1964).
- [9] Berkowitz, R. J. Acquisition and transmission of mutans streptococci. *Journal of the California Dental Association* **31**, 135–8 (2003).
- [10] Hudson, R. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337–338 (2001).
- [11] Okinaga, T., Niu, G., Xie, Z., Qi, F. & Merritt, J. The hdrRM operon of *Streptococcus mutans* encodes a novel regulatory system for coordinated competence development and bacteriocin production. *Journal of Bacteriology* **192**, 1844–1852 (2010).
- [12] Merritt, J., Zheng, L., Shi, W. & Qi, F. Genetic characterization of the hdrRM operon: a novel high-cell-density-responsive regulator in *Streptococcus mutans*. *Microbiology (Reading, England)* **153**, 2765–2773 (2007).
- [13] Bitoun, J. P., Nguyen, A. H., Fan, Y., Burne, R. A. & Wen, Z. T. Transcriptional repressor Rex is involved in regulation of oxidative stress response and biofilm formation by *Streptococcus mutans*. *FEMS Microbiol Lett* **320**, 110–7 (2011).
- [14] He, X. *et al.* The cia operon of *Streptococcus mutans* encodes a unique component required for calcium-mediated autoregulation. *Mol Microbiol* **70**, 112–26 (2008).
- [15] Wu, C. *et al.* Regulation of ciaXRH operon expression and identification of the CiaR regulon in *Streptococcus mutans*. *J Bacteriol* **192**, 4669–79 (2010).
- [16] Klein, M. I. *et al.* Structural and molecular basis of the role of starch and sucrose in *Streptococcus mutans* biofilm development. *Appl Environ Microbiol* **75**, 837–41 (2009).
- [17] Klein, M. I. *et al.* *Streptococcus mutans* Protein Synthesis during Mixed-Species Biofilm Development by High-Throughput Quantitative Proteomics. *PLoS One* **7**, e45795 (2012).
- [18] Chattoraj, P., Mohapatra, S. S., Rao, J. L. U. M. & Biswas, I. Regulation of transcription by SMU.1349, a TetR family regulator, in *Streptococcus mutans*. *J Bacteriol* **193**, 6605–13 (2011).
- [19] Klein, M. I. *et al.* Dynamics of *Streptococcus mutans* transcriptome in response to starch and sucrose during biofilm development. *PLoS One* **5**, e13478 (2010). Klein, Marlise I DeBaz, Lena Agidi, Senyo Lee, Herbert Xie, Gary Lin, Amy H-M Hamaker, Bruce R Lemos, Jose A Koo, Hyun Y1-DE-6006-02/DE/NIDCR NIH HHS/United States Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't United States PloS one PLoS One. 2010 Oct 19;5(10):e13478.

- [20] Lemos, J. A. C., Abrantes, J. & Burne, R. A. Responses of cariogenic streptococci to environmental stresses. *Curr Issues Mol Biol* **7**, 95–107 (2005).
- [21] Matsui, R. & Cvitkovitch, D. Acid tolerance mechanisms utilized by *Streptococcus mutans*. *Future Microbiol* **5**, 403–17 (2010).
- [22] Kobayashi, H., Saito, H. & Kakegawa, T. Bacterial strategies to inhabit acidic environments. *J Gen Appl Microbiol* **46**, 235–243 (2000).
- [23] Liu, M. & Taylor, D. E. Characterization of gram-positive tellurite resistance encoded by the *Streptococcus pneumoniae* *tehB* gene. *FEMS Microbiol Lett* **174**, 385–92 (1999).
- [24] Tanzer, J. M., Börjesson, A. C., Laskowski, L., Kurasz, A. B. & Testa, M. Glucose-sucrose-potassium tellurite-bacitracin agar, an alternative to mitis salivarius-bacitracin agar for enumeration of *Streptococcus mutans*. *J Clin Microbiol* **20**, 653–9 (1984).
- [25] Black, C., Allan, I., Ford, S. K., Wilson, M. & McNab, R. Biofilm-specific surface properties and protein expression in oral *Streptococcus sanguis*. *Arch Oral Biol* **49**, 295–304 (2004).
- [26] Rolerson, E. e. a. The SloR/Dlg metalloregulator modulates *Streptococcus mutans* virulence gene expression. *Journal of Bacteriology* **188**, 5033–5044 (2006).
- [27] Okinaga, T., Niu, G., Xie, Z., Qi, F. & Merritt, J. The *hdrRM* operon of *Streptococcus mutans* encodes a novel regulatory system for coordinated competence development and bacteriocin production. *J Bacteriol* **192**, 1844–52 (2010).
- [28] Okinaga, T., Xie, Z., Niu, G., Qi, F. & Merritt, J. Examination of the *hdrRM* regulon yields insight into the competence system of *Streptococcus mutans*. *Mol Oral Microbiol* **25**, 165–77 (2010).
- [29] Chong, P., Drake, L. & Biswas, I. Modulation of *covR* expression in *Streptococcus mutans* UA159. *J Bacteriol* **190**, 4478–88 (2008).
- [30] Idone, V. *et al.* Effect of an orphan response regulator on *Streptococcus mutans* sucrose-dependent adherence and cariogenesis. *Infect Immun* **71**, 4351–60 (2003).
- [31] Bidossi, A. *et al.* A functional genomics approach to establish the complement of carbohydrate transporters in *Streptococcus pneumoniae*. *PLoS One* **7**, e33320 (2012).
- [32] Knutsen, E., Ween, O. & Håvarstein, L. S. Two separate quorum-sensing systems upregulate transcription of the same ABC transporter in *Streptococcus pneumoniae*. *J Bacteriol* **186**, 3078–85 (2004).
- [33] Kajimura, J. *et al.* O acetylation of the enterobacterial common antigen polysaccharide is catalyzed by the product of the *yiaH* gene of *Escherichia coli* K-12. *J Bacteriol* **188**, 7542–50 (2006).
- [34] Rodionov, D. A. *et al.* Transcriptional regulation of NAD metabolism in bacteria: NrtR family of Nudix-related regulators. *Nucleic Acids Res* **36**, 2047–59 (2008).
- [35] Li, L., Stoeckert, C. J., J. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research* **13**, 2178–89 (2003).